

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): BABA, et al.
Serial No.: Not yet assigned
Filed: January 30, 2004
Title: METHOD, APPARATUS, AND COMPUTER READABLE
MEDIUM FOR MANAGING MULTIPLE SYSTEM
Group: Not yet assigned

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

January 30, 2004

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s)
hereby claim(s) the right of priority based on Japanese Patent Application No.(s)
2003-057937, filed March 5, 2003.

A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP



Carl I. Brundage
Registration No. 29,621

CIB/alb
Attachment
(703) 312-6600

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 2003年 3月 5日
Date of Application:

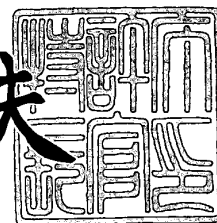
出願番号 特願2003-057937
Application Number:
[ST. 10/C]: [JP 2003-057937]

出願人 株式会社日立製作所
Applicant(s):

2004年 1月19日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫



出証番号 出証特2004-3000687

【書類名】 特許願

【整理番号】 K02014441A

【あて先】 特許庁長官殿

【国際特許分類】 G06F 12/00

【発明者】

【住所又は居所】 神奈川県横浜市戸塚区戸塚町 5 0 3 0 番地 株式会社日立製作所 ソフトウェア事業部内

【氏名】 馬場 恒彦

【発明者】

【住所又は居所】 神奈川県横浜市戸塚区戸塚町 5 0 3 0 番地 株式会社日立製作所 ソフトウェア事業部内

【氏名】 仲野 隆行

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

【代理人】

【識別番号】 100075096

【弁理士】

【氏名又は名称】 作田 康夫

【手数料の表示】

【予納台帳番号】 013088

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】

系切り替えシステムおよびその処理方法並びにその処理プログラム

【特許請求の範囲】

【請求項 1】

複数台から構成されるコンピュータシステムにおいて、現用系と現用系の処理を引き継ぐ待機系とを含むコンピュータシステムと、現用系と待機系で共有された少なくとも正ボリュームと副ボリュームからなる一対のボリュームと、前記一対のボリュームを対象とするボリュームレプリカ手段を有する一台または複数台のディスク装置と、現用系コンピュータシステムでのボリュームレプリカの実行の開始と完了とを待機系に通知する手段とを備えたことを特徴とする系切り替えシステム。

【請求項 2】

前記請求項 1 記載の系切り替えシステムにおいて、前記請求項 1 の通知手段によって送信されたボリュームレプリカ実行完了の通知を待機系コンピュータが受信し、副ボリューム情報反映処理を実行する手段を備えたことを特徴とする系切り替えシステム。

【請求項 3】

前記請求項 2 記載の系切り替えシステムにおいて、前記請求項 2 の手段の実行後に待機系が現用系に副ボリューム情報反映処理の実行完了を通知する手段を備えたことを特徴とする系切り替えシステム。

【請求項 4】

前記請求項 3 記載の系切り替えシステムにおいて、前記請求項 3 の通知手段によって送信された副ボリューム反映処理完了通知を現用系が受信する手段を持ち、現用系／待機系コンピュータシステム間で副ボリューム情報の一貫性を保証することを特徴とする系切り替えシステム。

【請求項 5】

前記請求項 4 記載の系切り替えシステムにおいて、ボリュームレプリカの実行と副ボリューム情報の一貫性の保証処理を実行中である現用系あるいは待機系コン

コンピュータシステムに障害が発生した場合に、該処理を継続して実行する手段を備えたことを特徴とする系切り替えシステム。

【請求項 6】

一台の現用系コンピュータシステムと、現用系に接続された少なくとも正ボリュームと副ボリュームからなる一対のボリュームと、前記一対のボリュームを対象とするボリュームレプリカ手段を有する一台または複数台のディスクから構成されるコンピュータシステムであって、少なくとも前記現用系の処理を引き継ぐ待機系コンピュータシステムが新たに追加されたことを契機として、現用系コンピュータシステムでのボリュームレプリカの実行状態を待機系に通信する手段と、前記実行状態から副ボリューム情報反映処理を実行する手段と、情報反映処理を実行後に現用系から待機系に情報反映処理完了を通知する手段と前記反映処理実行完了通知を現用系が受信する手段とを用いることにより、前記一対のボリュームを共有する現用系と待機系コンピュータとを含む高可用性コンピュータシステムとを備えたことを特徴とする系切り替えシステム。

【請求項 7】

前記請求項 7 の記載の系切り替えシステムにおいて、現用系あるいは待機系コンピュータシステムに障害が発生した場合に、ボリュームレプリカの実行と副ボリューム情報の一貫性の保証を継続して実行する手段とを備えたことを特徴とする系切り替えシステム。

【請求項 8】

現用系と現用系の処理を引き継ぐ待機系とを含み、現用系と待機系で共有された少なくとも正ボリュームと副ボリュームからなる一対のボリュームと、前記一対のボリュームを対象とするボリュームレプリカ手段を有する一台または複数台のディスク装置とをコンピュータシステム有するコンピュータシステムにおける系切り替え方法において、現用系コンピュータシステムでのボリュームレプリカの実行の開始と完了とを待機系に通知し、運用を開始することを特徴とする系切り替え方法。

【請求項 9】

現用系と現用系の処理を引き継ぐ待機系とを含み、現用系と待機系で共有され

た少なくとも正ボリュームと副ボリュームからなる一対のボリュームと、前記一対のボリュームを対象とするボリュームレプリカ手段を有する一台または複数台のディスク装置とをコンピュータシステム有するコンピュータシステムにおける系切り替えプログラムにおいて、現用系コンピュータシステムでのボリュームレプリカの実行の開始と完了とを待機系に通知し、運用を開始することを特徴とする系切り替えプログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は現用系と待機系を少なくとも有する障害許容性のあるコンピュータシステム技術に関する。

【0002】

【従来の技術】

現代社会において、コンピュータシステムは我々の生活を支える生活基盤として無くてはならない存在となっている。こうしたコンピュータシステムは24時間停止することなく、サービスを継続することが要求される。こうしたコンピュータシステムの一つとして、銀行などのオンラインシステムがあり、その中核処理としてデータベース業務がある。こうしたデータベース業務は絶えず更新されうするため、完全停止することが許されない。しかし、一方で扱うデータの保護のため、適宜バックアップを生成したいという需要がある。

データベースでは扱うデータを磁気ディスク装置に予め設定された記憶領域もしくは格納領域であるボリューム（以後、VOLと省略）に格納し、処理を行っている。VOLは上記記憶領域の単位であり、パーティション等の名称で呼ばれることもある。VOLは、物理VOL識別子（PVID）によって識別されており、このPVIDを利用してVOL情報を取得することによってコンピュータシステムから認識される。このPVIDとVOL情報は、ディスク管理プログラムによって取得され、オペレーティングシステム（OS）上のディスク管理情報バッファに保存される。データベースなどのアプリケーションやOSは、このバッファの情報を元にして認識されたVOLへのアクセス（読み書き）を行う。

【0003】

上述のように、データベースを完全停止させないでVOLの複製を実現する技術が重要である。ここでいう完全停止とは、システム障害があった場合でも業務上支障のない範囲の短時間に回復することを意図しており、システム利用者からシステムが停止しているというように見えないことをいう。このような技術として、特開2002-41368号に開示される技術のようなディスク内のデータをディスク装置内で複製（レプリカ）を実行するものがあり、これをVOLに適用したVOLレプリカという技術がある。このVOLレプリカには、データ複製元である正VOLとデータ複製先である副VOLとからなる一対のVOLを対象としたペア構成とペア分割という二つの手段からなる。ペア構成は、正／副VOLのPVIDとVOL情報とを含む全てのデータとを一致（同期）させることにより、正VOLの複製である副VOLを高速に生成する手段である。従って、ペア構成の状態では、正／副VOLのPVIDが一致した状態にあり、一つのVOLとして上位のコンピュータシステムに見える機能を提供する。一方、ペア分割は、ペア構成となっている一対のVOLのうち、副VOLのPVIDを正VOLとは違うPVIDに書き換える手段である。これにより、ペア構成時に上位のコンピュータシステムに対して一つに見えた一対のVOLを分割し、正VOLと副VOLとの別々のVOLとして見える機能を提供する。これらの二つの手段により、正VOLの複製を高速に生成し、複製した副VOLをコンピュータシステムから操作することができる機能が提供される。

【0004】

一方、短時間で回復する高信頼性を必要とするコンピュータシステムでは、処理を実行する現用系コンピュータシステムと、現用系に障害が発生した際に処理を引き継ぐ待機系コンピュータとを含む構成をとる。現用系で発生した障害の検出から待機系に処理を引き継ぐ手続きは、現用系と待機系を管理するクラスタプログラムによって提供される。処理を引き継ぐためには、アプリケーションやOSが使用しているデータを引き継ぐことが必要となる。例えば、上述のデータベースシステムでは、扱うデータが格納されているVOLに関する情報を引き継ぐ必要が生じる。

【0005】

しかし、上述のVOL複製技術は、ディスク装置内で複製を行うため、複製を実行したコンピュータ以外では複製したデータに関する情報をもたない。従って、クラスタプログラムを適用したコンピュータシステムにおいて、複製を実行したコンピュータで障害が発生し系切替が発生した場合、切替先のコンピュータは複製したデータに対してアクセスに失敗する。すなわち、系切替先が複製したVOLに対する処理を引き継ぐことができなくなってしまうという問題点がある。

【0006】

【特許文献1】

特開2002-41368号

【0007】

【発明が解決しようとする課題】

一方、短時間で回復する高信頼性を必要とするコンピュータシステムでは、処理を実行する現用系コンピュータシステムと、現用系に障害が発生した際に処理を引き継ぐ待機系コンピュータとを含む構成をとる。現用系で発生した障害の検出から待機系に処理を引き継ぐ手続きは、現用系と待機系を管理するクラスタプログラムによって提供される。処理を引き継ぐためには、アプリケーションやOSが使用しているデータを引き継ぐことが必要となる。例えば、上述のデータベースシステムでは、扱うデータが格納されているVOLに関する情報を引き継ぐ必要が生じる。

【0008】

しかし、上述のVOL複製技術は、ディスク装置内で複製を行うため、複製を実行したコンピュータ以外では複製したデータに関する情報をもたない。従って、クラスタプログラムを適用したコンピュータシステムにおいて、複製を実行したコンピュータで障害が発生し系切替が発生した場合、切替先のコンピュータは複製したデータに対してアクセスに失敗する。すなわち、系切替先が複製したVOLに対する処理を引き継ぐことができなくなってしまうという問題点がある。

【0009】

すなわち、VOLレプリカ（ペア構成・ペア分割）の対象となる正／副VOLを共有する現用系／待機系コンピュータシステムで障害が発生し、待機系が現用系の処

理を引き継いだ場合に、副VOLへのアクセスに失敗する。なぜなら、VOLレプリカによって、副VOLのPVIDが書き換えられ、この変更は現用系には反映されるが、一方の待機系ではディスク管理情報バッファに格納されたVOLレプリカを実行する以前の情報により、副VOLに対するアクセスしようとするためである。

【0010】

このように従来手法は、高信頼性のためのクラスタプログラムとVOLレプリカとの二つの手段を利用する場合には、現用系と待機系の処理を引き継ぐことのできない状況が発生するという問題があった。

【0011】

本発明の第一の目的は、VOLレプリカ手段による副VOLの変更を待機系に反映する方法及びシステムを提供することにある。

【0012】

本発明の第二の目的は、VOLレプリカ手段による副VOLの変更及び変更の反映を実施中に現用系あるいは待機系に障害が発生した場合に、この副VOLの変更及び変更の反映を引き継ぐ方法及びシステムを提供することにある。

【0013】

本発明の第三の目的は、VOLレプリカ手段実施後に現用系に障害が発生した場合に、現用系が実行していた処理を待機系に引き継ぐ方法及びシステムを提供することにある。

【0014】

本発明の第四の目的は、VOLレプリカ手段を利用しているコンピュータシステムを現用系とする待機系が新たに追加された場合に、待機系に副VOLの変更を反映する方法及びシステムを提供することにある。

【0015】

本発明の第五の目的は、VOLレプリカ手段を利用しているコンピュータシステムを現用系とする待機系が新たに追加され、待機系への副VOL変更の反映中に現用系あるいは待機系に障害が発生した場合に、この副VOLの変更及び変更の反映を引き継ぐ方法及びシステムを提供することにある。

【0016】

本発明の第六の目的は、VOLレプリカ手段を利用しているコンピュータシステムを現用系とする待機系が新たに追加後に現用系に障害が発生した場合に、現用系が実行していた処理を待機系に引き継ぐ方法及びシステムを提供することにある。

【0017】

【課題を解決するための手段】

本発明は、現用／待機系コンピュータシステムから構成される高可用性コンピュータシステムであって、現用／待機系がVOLレプリカの実行対象となる一対のVOLを共有しているコンピュータシステムにおいて、クラスタプログラムの起動時に、VOLレプリカによって変更される副VOLの物理名を取得する。例えば、VOLレプリカの設定ファイルに保存されている副VOLの物理名を読み込むことで、その物理名を取得する。

さらに、現用系でVOLレプリカの実行を行う際には、VOLレプリカの実行開始時と完了後に待機系に対して通知を行い、これにより待機系は現用系が副VOLを変更したことを認識する。

【0018】

待機系は副VOLの変更が行われたことを通知されると、副VOL状態の変更を反映する処理を行う。例えば、VOLレプリカでペア分割が行われた場合には、取得済みの副VOLの物理名を用いて、副VOLのPVIDが新たに設定されたPVIDの取得を行い、さらに、このPVIDを用いて副VOLの情報を取得する。これにより、変更後の副VOL情報が反映され、待機系が副VOLのアクセスを行うことができる。

【0019】

待機系は、副VOL情報の反映を完了すると、現用系に反映が完了したことを通知する。現用系はこの通知を受信することにより、副VOL情報の一貫性が保証されたことを認識する。

【0020】

各コンピュータシステムはVOLレプリカが現用系で実行されたかどうかと、副VOL情報が待機系に反映されているかどうかを表すVOLレプリカ状態を状態フラグとして所有する。また、状態ファイルとしてコンピュータシステム上に記憶して

おく。さらに、現用系／待機系で情報をやり取りする場合にこの状態フラグを相手側に通知することで現用系／待機系の処理状態を認識することができる。

【0021】

さらに、クラスタプログラムの起動時には、起動時のVOLレプリカ状態についても取得し、副VOL情報が現用／待機系で一致しているかを調査する。例えば、VOLレプリカ状態ファイルに保存されている副VOL反映状態を読み込むことで、副VOL情報が現用／待機系の両方に反映されているかを判断する。副VOL情報が待機系に反映されていない場合には、現用系が待機系に副VOL情報を反映するように通知し、待機系は副VOL情報の反映を行う。これにより、VOLレプリカ処理や副VOL情報の反映処理が中断された場合にも、その中断された処理を引き継いで再開することができる。

【0022】

【発明の実施の形態】

本発明に関する図と説明は、本発明を鮮明に理解するために適当な要素を示すために簡略化されており、既知の要素については省略していることを理解されたい。本技術中で従来技術の中には、本発明を実装するために他の要素が望ましく、かつ／または、必要とされると思われるものが幾つかある。しかし、技術中のこれらの要素は既知であり、本発明の理解を容易にするものではないので、ここでは説明しない。以下では、添付の図に関して詳細に説明していく。

【0023】

本実施例は現用系コンピュータで実行したVOLレプリカ対象となるVOLの情報を待機系コンピュータに反映するVOL情報一貫性保証システムを提供しようとするものである。

図1は、本実施例による現用／待機系コンピュータモデルのブロック図を表す。

図2は、本実施例によって解決される問題点をもつ従来手法の現用／待機系コンピュータモデルのブロック図を表す。

【0024】

図1、図2は、処理を行うコンピュータ層と処理に必要なデータを保存するデ

スク層から構成される。コンピュータ層は複数の現用系コンピュータ10と複数の待機系コンピュータ20からなる。各コンピュータは通信し合う手段01をもつものとする。また、各コンピュータ10、20は4つのプログラムを含み、以下の

- (1) コンピュータの動作を制御するオペレーティングシステム (OS) 11、21、
- (2) ディスク管理を行うディスク管理プログラム12、22、
- (3) 系切替による高可用性システムを実現するクラスタプログラム13、23、
- (4) アプリケーション14、24

である。

【0025】

前記クラスタプログラム13、23は前記通信手段01を用い、互いの情報を交換する機能や障害監視機能とを備える。一方、ディスク層は上位コンピュータ層によって共有されるディスク30を含む。前記ディスク30は二つの要素から構成される：(1) 情報を保存するVOL32、35、(2) ディスク装置内のVOLを制御するVOL管理機構31である。また、前記VOL32、35は上位の前記コンピュータ層10、20からのアクセス対象を識別するためのPVID33、36とVOL情報34、37とをそれぞれ有する。

【0026】

図2における手法では、現用系10の処理でディスク30が行ったVOLレプリカによるディスク上の副VOL情報の変更(1)は待機系20には反映されていないため、現用系10に障害が発生し(2)、クラスタプログラム13、23により系切替が生じた(3)場合に副VOL35に対するアクセスを行うことができない(4)。この結果、現用系10上のアプリケーション14の副VOL35に対する処理を待機系20上のアプリケーション24が正常に引き継ぐことができないことがある。

【0027】

図1で、本実施例では現用系10の処理でディスク30がVOLレプリカによるディスク上の副VOL情報の変更(1)を実行したことを契機として、通信手段01によって現用系クラスタプログラム13から待機系クラスタプログラム23に変更が行われたことを通知する(2)。これにより、副VOL35の制御権を一時的に現用系10から待機系20に切り替え、この変更された副VOL情報36、37を待機系20に反映させ

る(3)。反映後は、現用系10は副VOL35の制御権を再び獲得し、副VOL35に対する処理を実行する。これにより、副VOL情報が現用系／待機系コンピュータ間で一貫していることが保証されるため、この後、障害が発生した場合に副VOLに対する処理を引き継ぐことが可能である。

【0028】

図3は、本実施例における現用／待機系コンピュータのシステムブロックを簡易に示したものである。図3のシステムは一般に2つの要素から構成される：(1) 複数のアプリケーションコンピュータを含むコンピュータ層（前記10、20に対応）、(2) コンピュータ層が共有するデータを保存するディスク層（前記30に対応）である。図3では、説明を分かりやすくするために、各プログラムのラベルとして3桁の数字を用いている。また、数字は現用系コンピュータと待機系コンピュータは同一のプログラムに対して同じ下二桁の数字を用い、百の位は現用系コンピュータで1を、待機系コンピュータでは2で表している。以下では、先に各プログラムについて説明する。この説明では、各コンピュータのプログラムは現用系コンピュータ上のプログラム番号で説明しているが、待機系コンピュータ上の対応したプログラムの説明も兼ねる。

【0029】

ディスク300は、ボリューム管理部310と、VOLレプリカの対象となる正VOL320と副VOL330とを含み、前記正／副VOL（320、330）は、そのVOLの認識にPVID（321、331）と、VOLのアクセスに必要なVOL情報（332、333）とを含む。

前記ボリューム管理部310は、VOLレプリカを実行し、前記正VOL320と副VOL330のPVID（321、331）とVOL情報（332、333）とを変更する機能を有する。

現用系コンピュータ100は、OS130とクラスタプログラム120、ディスク管理プログラム150、アプリケーション150、及び、VOLレプリカ定義ファイル160、VOLレプリカ状態ファイル140を含む。

【0030】

前記OS130は、ディスク管理情報バッファ131を含む。また、前記ディスク管理プログラム150から前記ディスク300へのアクセスを仲介する。このアクセスでは、アクセスの結果を前記ディスク管理情報バッファ131に保存したり、前記ディ

スク300へのアクセスをせずに前記バッファ131に保存された情報を利用したりすることもある。前記VOLレプリカ定義ファイル160は、VOLレプリカの実行に必要な定義を含み、例えば、レプリカ対象となる前記正VOL320と副VOL330の物理名を含む。

【0031】

前記ディスク管理プログラム150は、現用系コンピュータ100上で実行されるディスク300をアクセスするプログラムを含み、例えば、VOLのロック制御プログラムと、VOLのPVIDやVOL情報を取得するプログラムと、VOLレプリカ実行プログラムとを含む。これらのプログラムの実行時には、ボリューム管理部310に指示を行ったり、前記ディスク管理情報バッファ131を読み込んだりすることもある。また、VOLレプリカ実行プログラムでは、前記VOLレプリカ定義ファイル160を利用することもある。

【0032】

クラスタプログラム120は、VOLレプリカの対象となる副VOLの識別情報を保存する副VOL識別情報バッファ121と、VOLレプリカの実行状態を保持するVOLレプリカ状態バッファ122と、他系との通信を行う通信部123と、他系及び自系の状態を監視する機能を提供する監視部124と、系切替に関する処理を行う系切替部125とを持つ。前記副VOL識別情報バッファ121は、前記VOLレプリカ定義ファイル160より読出した副VOLの識別情報を、保持しておくバッファである。

【0033】

前記監視部124は、前記アプリケーション110を監視することで自系の障害とMR CF（複製作成処理）の実行を検知する機能と、前記通信部125を介し待機系のクラスタプログラム220の通信部223と通信することで自系状態を通知する機能と他系状態の障害及びVOLレプリカの状態を検知する機能とを有する。

【0034】

前記系切替部125は、前記監視部124からの自系／他系の障害検知によって実行系・待機系の切替を行う系切替機能を有する。また、同切替部125は、同監視部124からの自系／他系のVOLレプリカ実行状態の検知によってディスク管理プログラム150を介して、VOLレプリカの実行を制御する機能と、その状態を前記VOLレ

プリカ状態フラグ122と前記VOLレプリカ状態ファイル140に保持する機能と、VOLレプリカを利用するアプリケーション110に副VOLの利用停止・再開の通知を行う機能とを有する。さらに、同切替部125は、前記VOLレプリカ定義ファイル160を読み込み、前記副VOL330を識別するために必要となる情報を前記副VOL識別情報バッファ121に保持する機能も有する。

【0035】

図4以降は、処理の流れを表している。各図中では、図3の数字と混同を避けるため、4桁の数字を用いている。図4～図12中の数字は、上二桁がそれぞれの図番号に対応した数字であり、下二桁のうち、01～20までを現用系コンピュータシステム上の処理、21～40までを待機系コンピュータの処理、41～60までをディスク装置の処理を示している。また、各コンピュータシステムとディスク装置とで行われるデータ交換処理を80～99で示している。また、以下の説明で、現用系の処理を用いて説明している場合であっても、特に明記のない場合は、その処理に対応する待機系の処理も同様に行われることを表すこともある。

【0036】

図4、図5は、本実施例による現用／待機系コンピュータモデルの処理の流れを簡易に表したものであり、図4はVOLレプリカ手段のうち、ペア分割を実行した場合、図5はペア再構成を実行した場合の処理を表す。

【0037】

図4、図5で、処理の流れは大きく二つの段階に分けられる：(1) 副VOL情報反映処理を行う上で必要となる情報をVOLレプリカ実施前に処理しておく前処理部、(2) 現用系でのVOLレプリカ手段実行を含む副VOL情報一貫性保証処理部である。各処理の詳細について、図3のシステムブロック図と対応づけて、以下に順に説明する。

【0038】

まず、前記前処理は図4、図5で共通する処理であり、また現用系と待機系コンピュータでも共通する処理を行う。前記前処理は第一にVOLレプリカ定義の読み込み0401を行う。これは、現用系コンピュータ100上のクラスタプログラム130がVOLレプリカ定義ファイル160を読み込む処理を表す。この定義ファイルから、VO

Lレプリカの対象となる副VOLの物理名を取得する（副VOL物理名取得処理0402）。ここまでで前処理が終了し、現用系100は副VOL変更が行われるまで本実施例が適用されていない場合の処理を実行し、通常現用状態0403になる。

【0039】

次に、副VOL変更が開始されると、前記一貫性保証処理部が実行される。一貫性保証処理は全部で三つの段階に分けられる。(1) Stage X: 現用系でVOLレプリカを実行する副VOL変更処理、(2) Stage Y: Stage Xで変更された副VOL情報を待機系に反映させる副VOL変更反映処理、(3) Stage Z: 副VOLに対する運用を再開する副VOL運用再開処理である。

【0040】

また、前処理は図4、図5で共通する処理であるが、副VOL情報一貫性保証処理は、Stage Xでの副VOL情報の変更内容が異なり、Stage Y、Stage Zでの副VOL情報反映処理で副VOLロックと副VOLへのアクセスが必要であるかが異なるため、図4、図5で別の処理が行われる。

【0041】

以下で順にStage毎の処理の流れを説明する。

Stage Xにおいて、まずディスク管理プログラム150によって、副VOL変更が実行されると、クラスタプログラム120上の系切替部125に通知される（副VOL変更通知0404）。この通知を受け、前記系切替部125は監視部124、通信部123を介し、待機系200のクラスタプログラム220に副VOL変更を実行することを通知する（矢印0581右向き）。待機系200は、前記通知0581を前記クラスタプログラム220上の通信部223、監視部224を介して、系切替部225で受信し、現用系100で副VOL変更処理が実行されることを認識する（副VOL変更開始認識処理0524）。認識後、現用系100から待機系200に対して実行を通知したのと逆の経路によって、待機系200は認識したことを現用系100に通知し（矢印0581左向き）、副VOL変更が完了されるのが通知されるまで待つ。

【0042】

次に、現用系100で前記系切替部125が前記待機系の認識通知0581を受信すると、前記ディスク管理プログラム130に処理を戻し、副VOL変更を伴う処理を実施す

る（VOLレプリカ実施0405）。この処理は、OS130あるいはOS上のディスク管理情報バッファ131を介して、ディスク装置300上のボリューム管理部310に処理が渡される。前記管理部310は、副VOLの制御権を獲得し、以降の処理を実行する。まず、図5のペア構成0541の場合、前記管理部310は副VOL330のPVID331が正VOL330のPVID321と同じ値に変更し、処理が終了する（矢印0582）。一方、図4のペア分割の場合は、前記管理部310が前記ペア構成0541により正VOLのPVID321と同じ値になっている前記副VOLのPVID331の値をユニークな別の値へと、また副VOLのVOL情報332もペア構成を行われる前の値から変更し、処理が終了する（矢印0482）。

以上がStage Xの処理であり、これによって現用系で副VOL変更処理を実施されることを待機系に通知することができる効果をもたらす。これは、待機系がこの副VOL変更完了を認識する前に現用系に障害が発生した場合に、前記変更処理が実施されたかどうかを待機系が認識できるようという効果をもたらし、これによって障害発生後に引き継ぐべき処理を認識することができるようになる。

【0043】

Stage Xの処理に続いて、Stage Y以降が実行されるが、ペア分割時（図4）とペア構成時（図5）で異なる処理になる。以下に図4、図5の順に詳細な処理を示す。

ペア分割時（図4）の場合、待機系200が前記副VOL320に対して処理を実行できるように、現用系100は前記VOLレプリカ処理0405時に獲得された副VOL320の制御権を解放する（副VOL解放0406）。この処理は、前記系切替部125が前記管理プログラム130を介して、前記VOLレプリカ処理0405で獲得した副VOLの制御権を解放するように前記管理部310が呼ばれる（矢印0483右向き）。前記管理部310は、副VOL320の制御権を解放する（副VOL解放0442）ことで完了する。

【0044】

前記副VOL解放処理0406が終了すると、副VOL変更完了通知処理0407が実行され、現用系100の系切替部125は、実行を通知した時と同様のパスを通じて、待機系200上の系切替部225に対して副VOL変更が完了したことを通知する（矢印0484右向き）。現用系100は、前記副VOL変更通知を待機系が受信したことを確認して副

VOL変更通知処理0407を終了し、待機系が変更された副VOL情報の反映を完了するのを待つ。

【 0 0 4 5 】

一方、前記副VOL変更通知を受けた待機系200上の系切替部225は、現用系で副VOL変更が行われたことを認識し（副VOL変更完了認識処理0425）、ディスク管理プログラム230を介して、副VOL情報の反映処理を行う。まず、副VOLの制御権の獲得を行う（副VOL獲得0426）。副VOLの獲得処理は前記副VOLの解放処理と同様の処理の流れで行われる（矢印0485、副VOL獲得0443）。

【 0 0 4 6 】

前記副VOL獲得処理0426の後、副VOLのPVID取得処理0427が行われる。前記PVID取得処理0427では、前記系切替部225が前記前処理0422で取得した前記副VOL物理名を用い、ディスク管理プログラム0230を実行する。前記プログラム0230は OS上のディスク管理情報バッファ231を介さず、ディスク装置300上の前記ボリューム管理部310を直接呼び出し（矢印0486左向き）、前記管理部310は副VOL330のPVID331を読み出し（副VOL PVID取得0444）、返す（矢印0486右向き）。これにより、前記系切替部225は前記副VOLのPVID331を得る（副VOL PVID取得処理0427）。この時、前記ディスク管理プログラム250は、前記処理0427で取得した副VOLのPVID331をディスク管理情報バッファ231に格納する。

【 0 0 4 7 】

前記副VOL PVID取得処理0427の後、副VOL情報取得処理0428が行われる。前記VOL情報取得処理0428では、前記系切替部225が前記PVID取得処理0427で取得した副VOLのPVID331を用い、ディスク管理プログラム0230を実行する。前記プログラム320は、前記処理0427同様の処理の流れで前記管理部310を呼び、副VOLのVOL情報332を取得する（矢印0487、副VOL情報取得処理0445）。この時も前記処理0427と同様に、前記ディスク管理プログラム250は取得された前記副VOL情報332をディスク管理情報バッファ231に格納する。

【 0 0 4 8 】

前記処理0428の完了後、前記副VOL獲得処理0426と同じ処理の流れで、前記系切替部225は副VOLの解放処理を行う（副VOL解放処理0429、0446、矢印0488）。

一方、ペア構成の場合（図5）、副VOL320は変更され、正VOL310と併せて一つのVOLとして現用系100/待機系200から見なされる状態にある。そのため、待機系200がディスク管理情報バッファ上に格納している副VOLのPVID331、VOL情報332により、存在しない変更前の副VOLにアクセスしエラーを発生することを消去する必要がある。従って、現用系100は副VOL変更完了通知処理0506により待機系200に副VOL変更完了を通知する（矢印0583）。ここで前記処理0506、通知（矢印0583）、及び待機系の副VOL変更完了認識処理0525は、前記処理0407、通知（矢印0484）、処理0425と同等の処理が行われ、待機系200上の系切替部225に副VOLが変更されたことが認識される。

【0 0 4 9】

副VOL変更を認識した前記系切替部225は、ディスク管理プログラム230を介して、副VOL情報の反映処理である副VOL情報削除処理0526を実行する。前記処理0526では、前記プログラム230がディスク管理情報バッファ231に対して処理を行い、前期バッファ231上に格納されたペア分割状態における副VOLのPVID331及びVOL情報332を消去する。

【0 0 5 0】

以上で、副VOL情報反映処理Stage Yが完了する。これによって、現用系で実施されたVOLレプリカによって変更された副VOL情報が待機系に反映される、ペア分割時（図4）においては待機系が副VOLのアクセスを行うことが可能となる効果と、ペア構成時（図5）においては待機系が存在しない副VOLにアクセスを行うのを防ぐ効果とをもたらす。

【0 0 5 1】

ここで説明するペア構成時とは、レプリカの開始をいい、副VOLは副VOLのPVIDとVOL情報が変更され、上位層からは隠蔽されてみえる。そのため、上位層からは正VOLのみが認識可能であり、その結果、正VOLに対してのみアクセスが可能になる。但し、本装置では正VOLへのwriteは副VOLにも、同期または非同期で反映される。

【0 0 5 2】

ここで説明するペア分割時とは、レプリカの終了をいい、VOLの内容はペア分

割時の正VOLと同じ状態が保持されたまま、副VOLのPVIDとVOL情報が変更され、正VOLと分離した一つのVOLとして上位層から認識可能になる。上位層からは正VOLと副VOLと個別にアクセス要求を出すことができる状態となる。

【 0 0 5 3 】

Stage Yの後、Stage Zが実行される。待機系200は前記処理0429の後、反映完了通知0430を実行し、現用系200の系切替部225は、実行を通知した時と同様のパスを通じて、待機系100上の系切替部125に対して副VOL情報が待機系に反映されたことを通知する（矢印0489右向き）。現用系200は、前記通知を現用系が受信したことを確認して前記処理0430を終了する。

ここで、前記処理0410において、正／副VOL間で、PVIDやVOL情報以外の、データが一致していない場合がある。これは以下の理由による。正／副VOLのデータは一致していたペア分割（副VOL変更処理0401）の後、前記処理0410が実行されるまでに、正VOLの運用が継続され、その結果、正VOLが更新されている場合があるためである。ここで、正／副VOLの一致が必要である場合には、副VOL変更処理0401から前記処理0410までの間、正VOLの更新を一時的に制限することや、制限を行わない場合には、前記処理0410において、正／副VOLの運用を開始する前に、正／副VOLの同期化を行うこともある。

【 0 0 5 4 】

一方、前記通知（矢印0489）を受けた現用系の系切替部125は、待機系で副VOL情報が反映されたことを認識し（反映完了認識処理0408）、ディスク管理プログラム130を介して、副VOLの制御権の獲得を行う（副VOL獲得0409）。以上で、Stage Zが完了し、現用系は通常運用状態0410に移行し、正／副VOL両方の運用を開始し、待機系は通常待機状態0431に移行する。

【 0 0 5 5 】

一方、ペア構成時（図5）は、現用系100／待機系200は前記処理0430、前記情報反映完了通知（矢印0489）、前記処理0408と同様に、待機系の反映完了通知処理0527、情報反映完了通知0583、現用系の反映完了処理0507が実行される。以上で、Stage Zが完了し、現用系は通常運用状態0508に移行し、正VOLの運用を継続し、待機系は通常待機状態0528に移行する。

以上の一連の処理によって、VOLレプリカによって変更が加えられた副VOLのPV ID及びVOL情報を現用／待機系の両方が認識し、副VOLの情報について一貫性が保証される効果がもたらされる。これにより、以降、現用系で障害が発生しても、待機系が副VOLにアクセスすることができ、正常に正／副VOL双方の業務を引き継ぐことができるようになり、課題を解決することができる。

【0056】

図6～図12は、本実施例の処理の流れを詳細に表したフローチャートである。以下では、現用系と待機系の処理がある場合は、左側に現用系の処理を、右側に待機系の処理をあらわしている。また、図中の各処理のうち、図3～図5の処理と対応させて説明するが、説明を簡単にわかりやすくするため、図3～図5の説明にあって以下で説明をしていない場合もある。

【0057】

図6は、現用／待機系コンピュータが実行する前記前処理を詳細に表したフローチャートである。前処理では、現用系／待機系で同様の処理が行われるため、以下では現用系について説明する。

まず、VOLレプリカ定義ファイルの有無を確認する処理0601を行い、VOLレプリカ定義ファイル140にアクセスする（矢印0691）。存在しない場合は、VOLレプリカ処理が実行されないため、特に何の処理を行う必要がなく、前処理を終了する。

一方、存在する場合は、前記ファイル140を読み込み（処理0602、矢印0692）、VOLレプリカの対象となる副VOLの物理名の取得処理0603を行う。ここで、前記処理0602が前記処理0401に、前記処理0603が前記処理0402に対応する。

さらに、VOLレプリカ状態ファイル140よりVOLレプリカ状態を読み込み、VOLレプリカ状態フラグ122に格納する（処理0604、矢印0693）。これは次の理由による。VOLレプリカを実施中にクラスタプログラムを再起動した場合に、待機系に副VOL情報を反映する必要があるかを認識するためである。

【0058】

以上で、前処理は終了し、通常現用状態0404に移行する（図7、7A）。

図7は、従来のクラスタプログラムによる障害監視／系切替処理に対して、VOL

レプリカ／副VOL情報一貫性保証処理との関係を示すフローチャートである。図4、図5では、現用系の通常運用状態（処理0410、0508）と待機系の通常待機状態（処理0431、0528）の処理に対応している。

【 0 0 5 9 】

図7で、現用系100上のクラスタプログラム120はまず、系切替部125で自系状態をチェックし（処理0701）、系切替が必要かを判断する（処理0702）。前記自系状態チェック処理0701は、前記クラスタプログラム120とアプリケーション110との間で行われる通信（矢印0782）を含む。また、前記通信0782には系切替が必要となるアプリケーション障害、あるいはアプリケーションからVOLレプリカの実行要求があったかどうかといった情報が含まれる。

【 0 0 6 0 】

前記処理0702で、系切替が必要である場合は、前記クラスタプログラム120は前記アプリケーション110に系切替のために処理を中断するように通知を行い（矢印0783）、系切替部125により待機系に切り替わる（処理0703）。その後、元現用系だったコンピュータシステム100は待機系に切り替わったので、前記クラスタプログラム120は待機系の監視処理（図7、7B）を実行する。

【 0 0 6 1 】

一方、系切替が必要でない場合は、前記系切替部125と待機系200の系切替部225は監視部124、224と通信部123、223を介して通信を行い（矢印0781）、自系状態の通信と待機系の状態をチェックする（処理0704）。この時、前期通信0781は前記VOLレプリカ状態フラグ122に格納された自系のVOLレプリカ状態の交換を含む。これには次のような理由がある。自系／他系のVOLレプリカ状態を見ることで、両系で副VOL状態が変更されたか／反映されたかが認識できるために、VOLレプリカ中に系切替や待機系コンピュータの追加が生じた場合に副VOL情報の反映が必要であるかを判断することができるためである。また、以降の説明文中で現用／待機系間で、クラスタプログラム同士、あるいは系切替部同士が通信する記述がある場合、特に明記されていなくても前記通信0781と同様の処理が行われる。

【 0 0 6 2 】

続いて、前記系切替部125は副VOL情報反映が必要であるかの判断を行う（処理0705）。これには次のような理由がある。通常状態では副VOL情報変更処理は現用系で、副VOL情報反映処理は待機系で行われるが、系切替直後は今現用系であるコンピュータシステムは元待機系であり、副VOL情報反映処理を行わずに系切替が行われた可能性があるからである。前記処理0706で副VOL情報反映が必要である場合は、前記クラスタプログラム120は障害回復時の副VOL反映処理（図11、11A）を実行する。必要でない場合は、前記クラスタプログラム120は前記0782で確認したVOLレプリカ実行要求の有無によってVOLレプリカを実行する必要があるかを判断する（処理0706）。前記処理0706で実行する必要がある場合には、前記クラスタプログラム120はVOLレプリカ実行処理（図8、8A）を実行する。必要が無い場合は、前記クラスタプログラム120は再び前記自系チェック処理0701へと戻って処理を継続する。

【0063】

次に、待機系200上のクラスタプログラム220はまず前記処理0701同様の自系状態のチェックを行い（処理0721、矢印0783）、自系が正常であるかを判断する（処理0784）。前記処理0784で、自系が異常である場合には、前記クラスタプログラム220上の系切替部225は監視処理を停止し、終了する（処理0723）。一方、自系が正常である場合は、前記系切替部225は前記通信0781によって現用系100と通信を行う。続いて、前記クラスタプログラム220は副VOL情報反映が必要であるかの判断を行う（処理0725）。副VOL情報反映が必要である場合は、前記クラスタプログラム220は障害回復時の副VOL反映処理（図12、12B）を実行する。必要でない場合は、前記系切替部225は前記通信0781で得た現用系100の状態により、系切替が必要かを判断する（処理0726）。系切替が発生した場合は、前記クラスタプログラム220はアプリケーション210に通知を行い（矢印0784）、現用系へと切り替わる（処理0727）。さらに前記クラスタプログラム220は現用系の監視処理（図7、7A）の実行を行う。一方、系切替が必要でない場合は、前記クラスタプログラム220は前記通信0781で得た現用系でのVOLレプリカの実行の有無により、現用系100でVOLレプリカが実行されるかを判断する（処理0728）。実行される場合には、前記クラスタプログラム220はVOLレプリカ実行処理（図8、8B）を実行

し、必要が無い場合は、再び前記処理0721へと戻り処理を継続する。

【 0 0 6 4 】

図8、図9、図10は、図4、図5で示した前記副VOL情報一貫性保証処理を詳細に表したフローチャートであり、図8はStage X（副VOL変更処理）、図9はStage Y（副VOL変更反映処理）、図10はStage Z（副VOL運用再開処理）を表している。

【 0 0 6 5 】

図8で、現用系100はVOLレプリカを実行する前に、副VOL変更開始を待機系へ通知し（処理0801、矢印0881）、待機系200はこの通知を受信する（処理0821）。待機系200は前記処理021後、VOLレプリカ処理が実施中であることを示す状態フラグB1を設定し（処理0822）、副VOL状態反映処理（図9、9B）を実行する。ここで、前記フラグ設定処理0822は、系切替部225がVOLレプリカ状態フラグ222を介して、VOLレプリカファイル240に格納する処理（矢印0884）を含む。以降の説明でフラグ設定処理では特に明記しない場合でも、前記0884と同様の格納処理が行われる、

一方、現用系100は前記処理0801の後、前記クラスタプログラム120はVOLレプリカ実行開始を示す状態フラグA1を設定し（処理0802）、VOLレプリカを実行する（処理0803）。前記VOLレプリカ処理0803が終了すると、前記クラスタプログラム120は実行完了を示す状態フラグA2を設定し（処理0804、矢印0883）、副VOL状態反映処理（図9、9A）を実行する。

【 0 0 6 6 】

ここで前記処理0801～0802は前記処理0404または0504に、0803～0804は前記処理0405または0505に、前記処理0821～0822は前記処理0424または0524に、前記通信0881は前記通信0481または0581にそれぞれ対応する。

【 0 0 6 7 】

図9で、現用系100のクラスタプログラム120はまず、VOLレプリカ処理でペア分割が行われたかを判断する（処理0901）。このVOLレプリカ処理は前記0803または後述する処理1104で実施されるものである。この前記処理0901は次のような理由による。なぜなら、ペア分割時は、待機系が副VOLへのアクセスを行うため、副VOLの解放を行う必要があり、一方、ペア再構成時は、副VOLの解放が必要でな

いため、副VOLの運用を現用系で開始できるためである。従って、ペア分割時、前記クラスタプログラム120は副VOLの解放（処理0902）を行う。その後、前記クラスタプログラム120は待機系に副VOL変更完了を通知し（処理0903、矢印0981）、待機系200が副VOL状態反映処理を実行中であることを示す状態フラグA3を設定した（処理0904）後、副VOL運用再開処理（図10、10A）を実行する。一方、待機系200上のクラスタプログラム220は、前記通信0981を受信すると（処理0921）、副VOL変更反映処理の実行開始を示す状態フラグB2を設定する（処理0922）。その後、前記処理0901同様にペア分割が行われたかを判断する（処理0923）。これは次のような理由による。なぜなら、ペア分割時は待機系が副VOLへのアクセスを行うためロック制御を行う必要があり、一方、ペア構成時は副VOL情報を削除するだけで副VOLへのアクセスは不要であるからである。従って、ペア分割時、クラスタプログラム220は副VOLのPVIDを取得（処理0925）後、副VOLロックを獲得する（処理0926）。そして、副VOL情報の取得し（処理0927）、副VOLロックの解放を行う（処理0928）。一方、ペア再構成時、前記クラスタプログラム120は副VOLのPVIDを含む副VOL情報の削除を行う（処理0924）。これにより、それぞれの場合において待機系200への副VOL情報の反映が完了する。その後、反映完了を示す状態フラグB3を設定し（処理0929）、副VOL運用再開処理（図10、10B）を実行する。

【 0 0 6 8 】

ここで、前記処理0902は前記処理0406に、前記処理0903～0904は前記処理0407あるいは0506に、前記処理0921～0922は前記処理0425あるいは0525に、前記処理0924は前記処理0526に、前記通信0981は前記通信0484あるいは0583に、それぞれ対応する。さらに、前記処理0925～0928は順に前記処理0426～0429にそれぞれ対応する。

【 0 0 6 9 】

図10で、待機系200上のクラスタプログラム220は副VOL情報反映が終了したことを現用系100上のクラスタプログラム120に通知する（処理1021、矢印1081）。前記処理1021後は、前記クラスタプログラム220は情報が反映されたことを示す状態フラグ0を設定し（処理1022）、再び障害監視処理（図7、7B）に戻り処理を

続ける。一方、前記クラスタプログラム120は前記通信1081を受信すると（処理1001）、情報が反映されたことを示す状態フラグ0を設定する（処理1082）。この後、前記処理0901同様にペア分割が行われたかを判断する（処理1003）。これは次のような理由による。なぜなら、ペア分割時は、前記Stage Yにおいて副VOLの解放を行っており、再度副VOLの獲得を行う必要があり、一方、ペア分割時は、前記Stage Yにおいて副VOLの運用を現用系で開始済みであるためである。従って、ペア分割時の場合のみ、前記クラスタプログラム120は副VOLロックの獲得（処理1004）を実行した後、アプリケーション110に副VOL運用が可能であることを通知する（処理1005、矢印1083）。この後、再び障害監視処理（図7、7A）に戻り、処理を続ける。

【0070】

ここで、前記処理1001～1002は前記処理0408あるいは0507に、前記処理1004は前記処理0409に、前記処理1021～1022は前記処理0430あるいは0527に、前記通信1081は前記通信0489あるいは0583にそれぞれ対応する。また、前記処理1005は前記処理0410中に含まれる。

【0071】

図11は、障害発生時に現用系がVOLレプリカ処理と副VOLを引き継ぐ手続きを表す風呂チャートである。図11中の処理のうち、図8、図9、図10で説明した処理と同様の処理については、理解しやすくするために対応のみを示している。

まず、現用系100のクラスタプログラム120は最初にVOLレプリカ状態フラグ222を参照し、状態フラグが0であるかを判別する（処理1101）。前記処理1101で、0である場合には、副VOL情報は正しく反映されていることを意味する。そのため、この場合には、前記クラスタプログラム120は何も処理をせずに障害時の待機系への副VOL情報反映処理（図12、12A）を実行する。前記処理1101で、0で無い場合は、前記クラスタプログラム120は状態フラグがA1またはB1であるかを判別する（処理1102）。

【0072】

前記処理1102で、A1またはB1である場合にはVOLレプリカが実行中であって、まだ変更が現用系で完了していない状態を意味するので、前記クラスタプログラ

ム120は前記Stage X (図4) に対応するVOLレプリカを実行し、現用系に副VOL情報を反映する (処理1103~1105)。ここで前記処理1103~1105は前記処理0802~0804にそれぞれ対応する。一連の反映処理後は、待機系への副VOL情報反映処理・反映処理 (図12、12A) を実行する。

【0 0 7 3】

一方、前記処理1102で、A1またはA2で無い場合は、すでにVOLレプリカが実施されていることを意味している。そこで、実施後の待機系反映処理がどこまで処理されているかに応じて処理を行う。

【0 0 7 4】

まず、前記クラスタプログラム120は状態フラグがA2またはA3であるかを判別する (処理1106)。前記処理1106で、A2またはA3である場合には、現用系での副VOL情報は反映され、待機系での副VOL情報反映処理完了通知を待っている状態を意味するので、前記クラスタプログラム120は通知待ち状態を解除する (処理1107)。次に前記処理1106でA2でもA3でも無かった場合は、状態フラグがB2であるかを判別する (処理1108)。前記処理1108で状態フラグがB2であった場合には、情報反映処理中だった待機系が現用系に系切替した状態を意味するので、前記クラスタプログラム120は情報反映処理を継続し、副VOL情報を反映する。この反映処理は、前記Stage Y (図7、7B) と同様にペア分割かペア再構成かによって異なる処理が行われる (処理1109~1113)。ここで前記処理1109~1113は前記処理0923~0927にそれぞれ対応する。一方、前記処理1108で、B2ではない場合には状態フラグはB3であり、副VOL状態の反映が完了していた待機系が現用系に系切替した状態であることを意味するので、副VOL情報反映は完了しており、何もする必要がない。

【0 0 7 5】

以上によって、前記処理1102でA1またはB1でない場合にも副VOL情報の引き継ぎが完了したので、状態フラグを0に設定し (処理1114)、現用系のStage Z (図10) と同様にペア分割かペア再構成かによって副VOL運用開始処理が行われる (処理1115~1117、矢印1181)。ここで、前記処理1115~1117は前記処理1003~1005に、通信1181は通信1083にそれぞれ対応する。その後、待機系への副VOL情報

反映処理反映処理（図12、12A）を実行する。

【0076】

図12は、現用／待機系コンピュータが障害発生時に現用系コンピュータが引き継いだボリューム情報を待機系コンピュータに引き継ぐ手続きを表すフローチャートである。図12中の処理についても図11と同じように、図8、図9、図10で説明した処理と同様の処理については、理解しやすくするために対応のみを示している。

【0077】

まず、現用系100／待機系200のクラスタプログラム120、220はそれぞれ自系の状態フラグを他系に送信し、お互いの状態を認識する（処理1201、1221、矢印1281）。これは次のような理由による。待機系の状態フラグによって待機系の副VOL状態反映処理がすでに反映されているか（待機系状態フラグが0またはA2またはA3またはB3）を判断するためと、待機系で副VOL状態反映が必要な場合に現用系の状態フラグによって待機系が前記Stage X（現用系状態フラグがA1またはB1）を実行する必要がある場合を判断するためと、待機系の副VOL引継ぎ処理（Stage Y）を実行する場合のうち、現用系が副VOL状態を反映処理中の待機系が系切替した状態であり、副VOL状態の引継ぎが必要である（現用系状態フラグがB2）場合とで処理が異なるからである。

【0078】

従って、前記クラスタプログラム120、220は待機系状態フラグが0またはA2またはA3またはB3であるかを判別する（処理1202、1222）。前記処理1202、1222でYである場合には、現用／待機系は障害監視処理に戻って処理を継続する（図7、7A／7B）。一方、前記処理1202、1222で、Nである場合は待機系で副VOL状態の反映が必要なことを意味するため、前記クラスタプログラム120、220は現用系状態フラグがA1またはB1であるかを判断する（処理1203、1223）。前記処理1203、1222でA1またはB1である場合には、現用／待機系はStage Xから実行を行う（図8、8A／8B）。

【0079】

前記処理1203、1223で上記以外の場合は、待機系200はVOL反映処理を実行する

必要があるため、Stage Yを実行する（図9、9B）。一方、現用系100は、前記ク
ラストプログラム120で現用系状態フラグがB2であるかを判断する（処理1204）
。前記処理1204で、B2である場合には、前記副VOL状態の引継ぎが必要であるた
め、待機系のStage Y（図9、9B）と同様の処理を行い、副VOL情報を現用系に反
映する（処理1205～1208）。ここで前記処理1205～1207は前記処理0923～0925に
、前記処理1208は前記処理0927に対応する。前記処理1205～1208が完了すると、
現用系で情報反映が完了している状態になるため、前記処理0804と同様の処理で
状態フラグA2を設定する（処理1209）。そして、Stage Yを実行する（図9、9A
）。一方、前記処理1204にてB2で無い場合は、そのままStage Yを実行する（図9
、9A）。

以上の図11、図12に示される処理によって、図6～10で示される副VOL情報の一
貫性保証処理が行われている最中に障害が発生しても、継続して現用系が実行中
の処理を引き継いで副VOL情報の一貫性を保証する効果を得る。さらに上記の効
果と、図4、図5で示した一貫性保証処理後に障害が発生した場合に系切替が保証
される効果から、VOLレプリカを実行する現用系／待機系コンピュータシステム
で障害が発生してもVOLレプリカを含む処理が現用系から待機系に引き継ぐこと
が可能である高可用性システムを構築することができる効果をもたらす。

【0080】

以降、現用系で障害が発生しても、待機系が副VOLにアクセスすることができ
、正常に正／副VOL双方の業務を引き継ぐことができるようになり、課題を解決
することができる。

【0081】

なお、本実施例では、現用系／待機系に障害が発生した場合について示したが
、通信手段01により利用されるネットワークに障害が発生した場合にも、現用
系／待機系の優先度を考慮することで、本実施例の技術を適用することも可能で
ある。

【0082】

以上述べたように、副ボリュームの変更を生じるボリュームレプリカの実行を
契機として、待機系に変更された副ボリューム情報が反映されるので、ボリュー

ムレプリカ実行後に現用系に障害が発生した場合にも、待機系に処理を継続することができる。また、ボリュームレプリカや副ボリューム反映処理を実行中に障害が発生した場合も、障害発生時に行っていた処理を系切替後に引き継ぐことができる。

【0083】

【発明の効果】

以上述べたように、本発明によれば、VOLレプリカ手段による副VOLの変更を待機系に反映することが可能となる。

【図面の簡単な説明】

【図1】本実施例による現用／待機コンピュータシステムモデルの高位のシステムブロック図である。

【図2】現用／待機コンピュータシステムモデルを用いた従来の障害引継システムの高位ブロック図である。

【図3】本実施例によるコンピュータシステムの低位のブロック図である。

【図4】本実施例によるコンピュータシステムで現用系コンピュータがボリュームレプリカ処理のペア分割を実施した場合の現用／待機系コンピュータの手続きの概要を表すフローチャートである。

【図5】本実施例によるコンピュータシステムで現用系コンピュータがボリュームレプリカ処理のペア再構成を実施した場合の現用／待機系コンピュータの手続きの概要を表すフローチャートである。

【図6】現用／待機系コンピュータの前処理のフローチャートである。

【図7】現用／待機系コンピュータの障害監視／系切替処理のフローチャートである。

【図8】現用系コンピュータがボリュームレプリカ処理を実行完了する手続きを表す現用／待機系コンピュータのフローチャートである。

【図9】待機系コンピュータが変更された副ボリューム情報を反映させる処理を実行完了する手続きを表す現用／待機系コンピュータのフローチャートである

【図10】現用／待機系コンピュータで副ボリューム情報の反映後、障害監視／系切替処理に戻る手続きを表すフローチャートである。

【図 1 1】 障害発生時に現用系コンピュータがボリュームレプリカ処理と副ボリュームを引き継ぐ手続きを表すフローチャートである。

【図 1 2】 障害発生時に現用／待機系コンピュータが現用系コンピュータの引き継いだボリューム情報を待機系コンピュータに引き継ぐ手続きを表すフローチャートである。

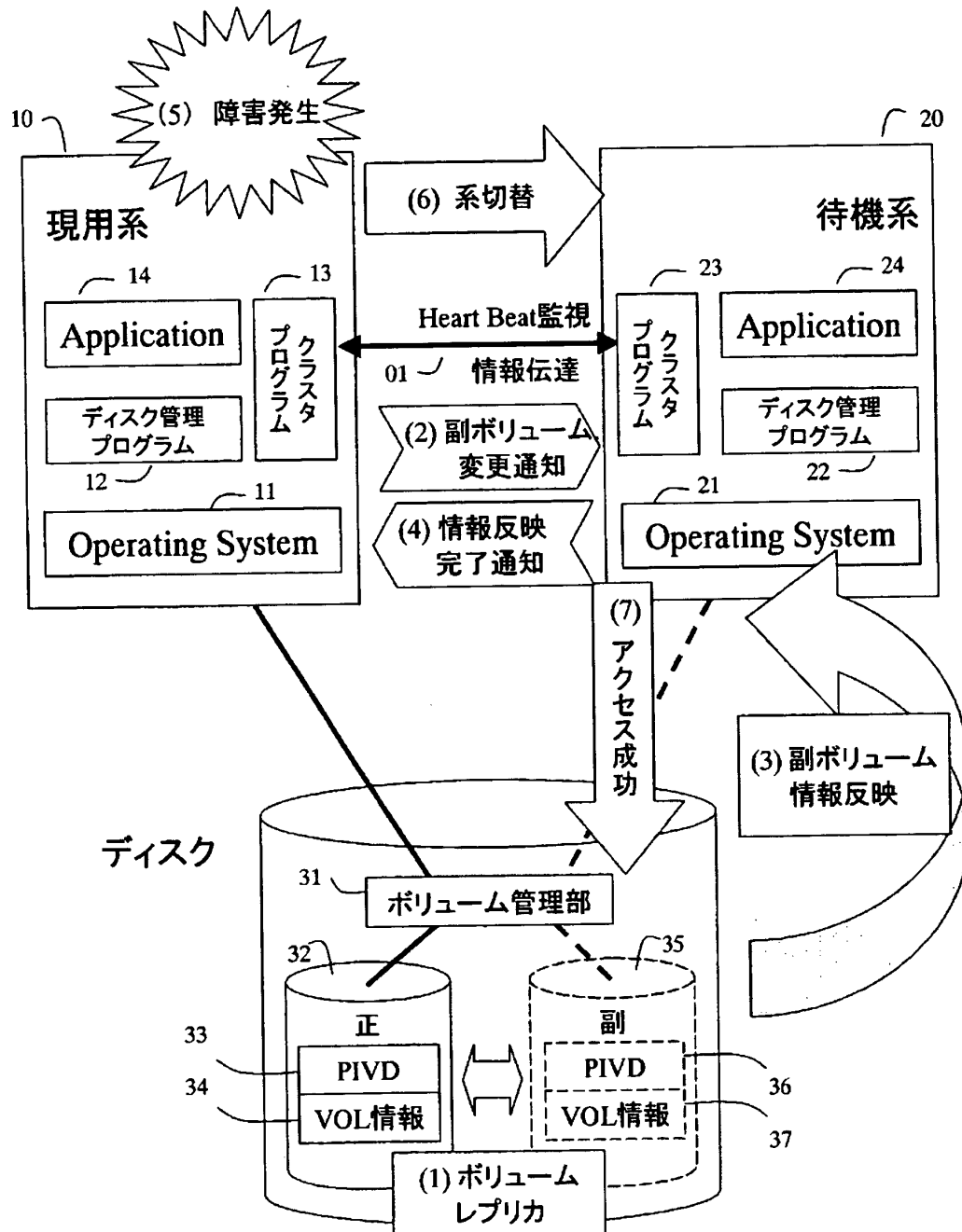
【符号の説明】

100・・・現用系コンピュータ
200・・・待機系コンピュータ
110、210・・・アプリケーション
120、220・・・クラスタプログラム
121、221・・・副ボリューム識別情報バッファ
122、222・・・ボリュームレプリカ状態フラグ
123、223・・・通信部
124、224・・・監視部
125、225・・・系切替部
130、230・・・オペレーティングシステム
131、231・・・ディスク管理情報バッファ
140、240・・・ボリュームレプリカ状態ファイル
150、250・・・ディスク管理プログラム
160、270・・・ボリュームレプリカ定義ファイル
300・・・ディスク
310・・・ボリューム管理部
320・・・正ボリューム
330・・・副ボリューム
321、331・・・物理ボリューム識別情報
322、332・・・ボリューム情報

【書類名】 図面

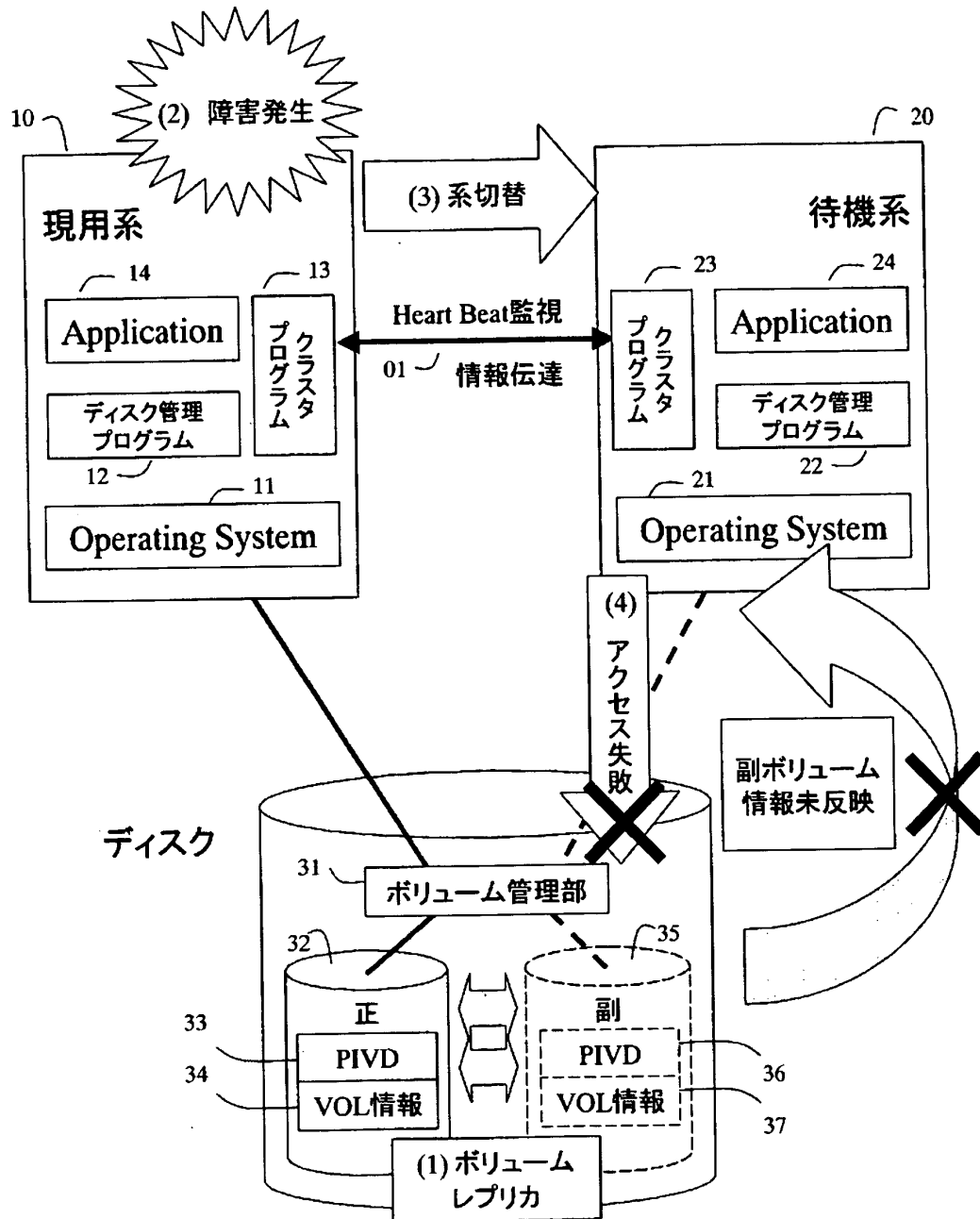
【図 1】

図1



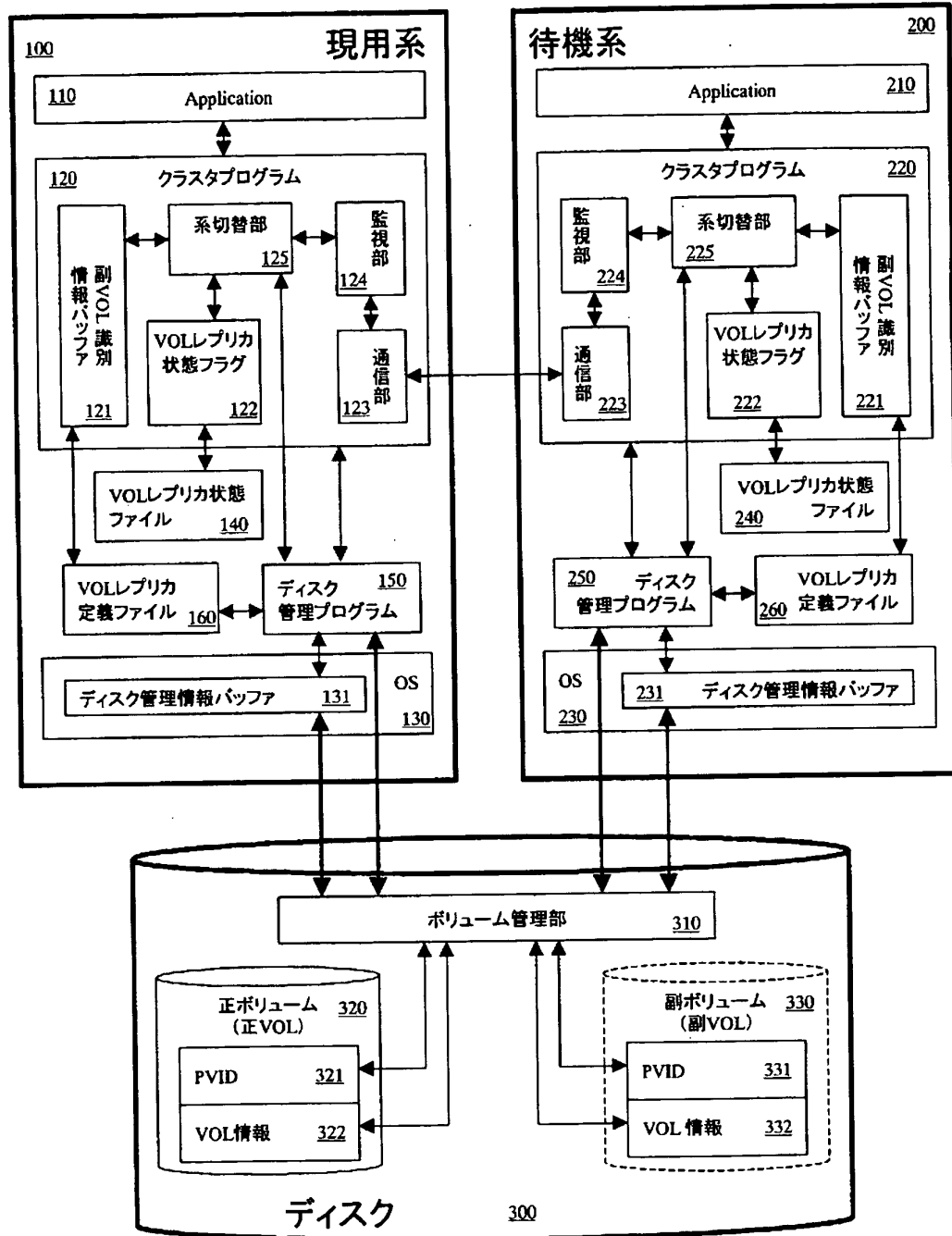
【図2】

図2

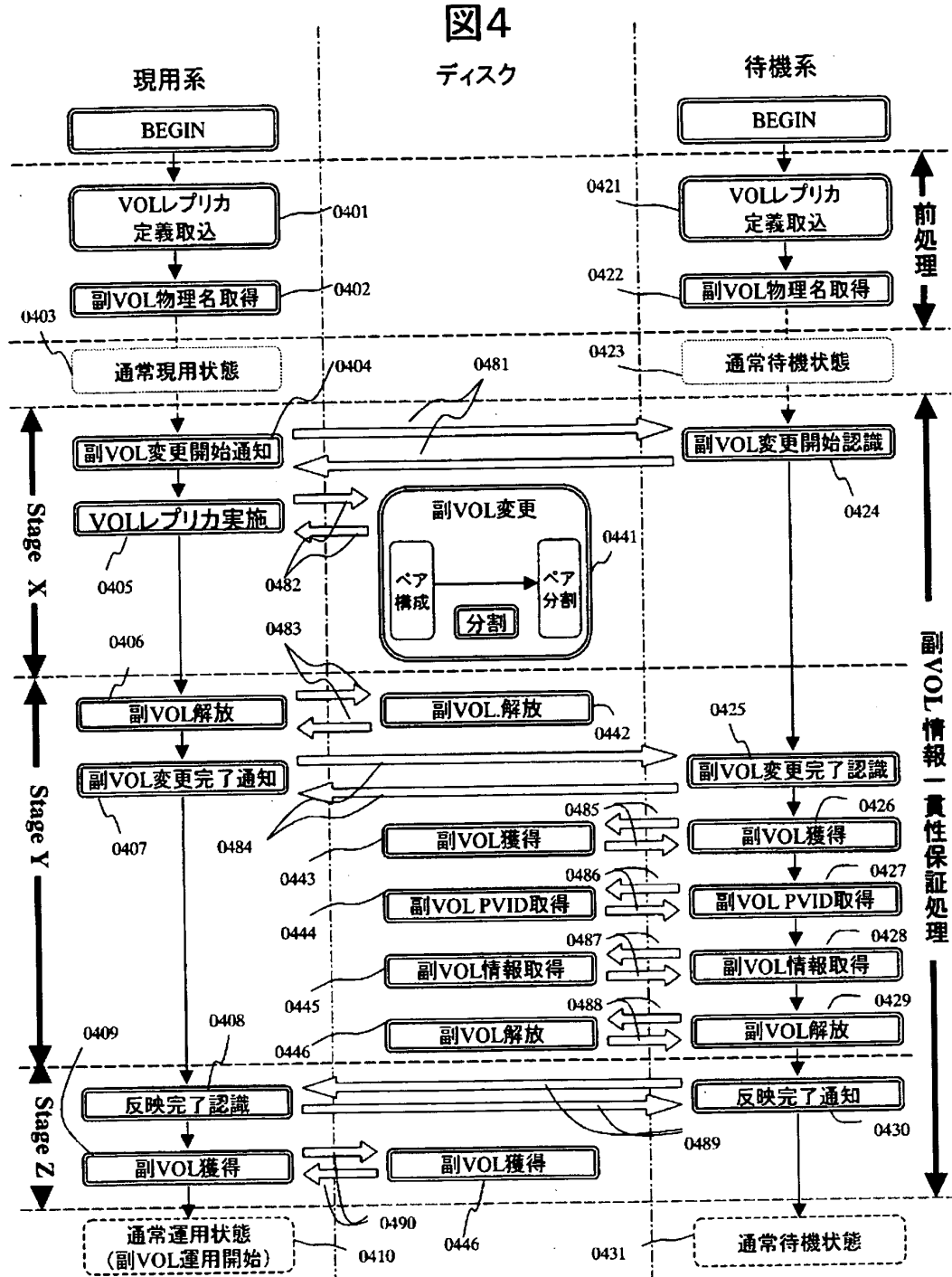


【図 3】

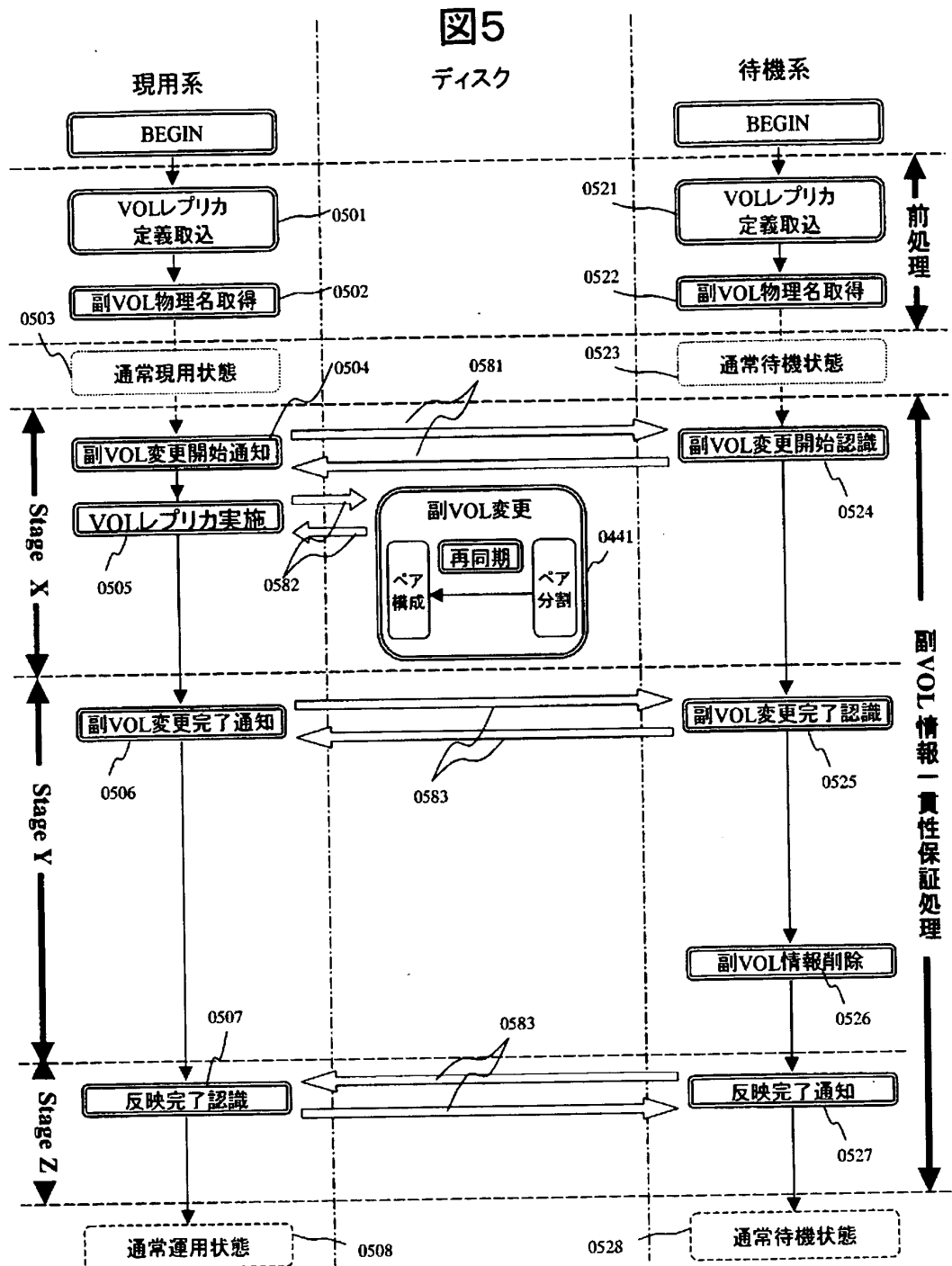
図3



【図 4】

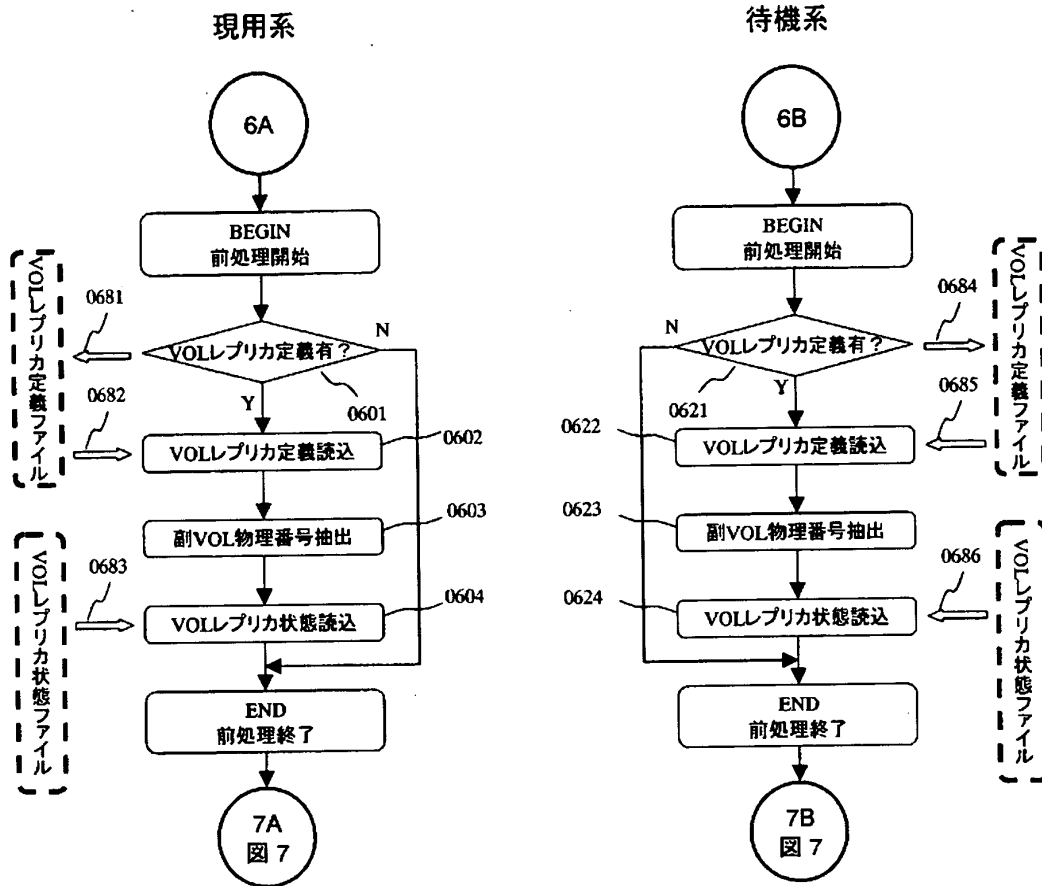


【図5】



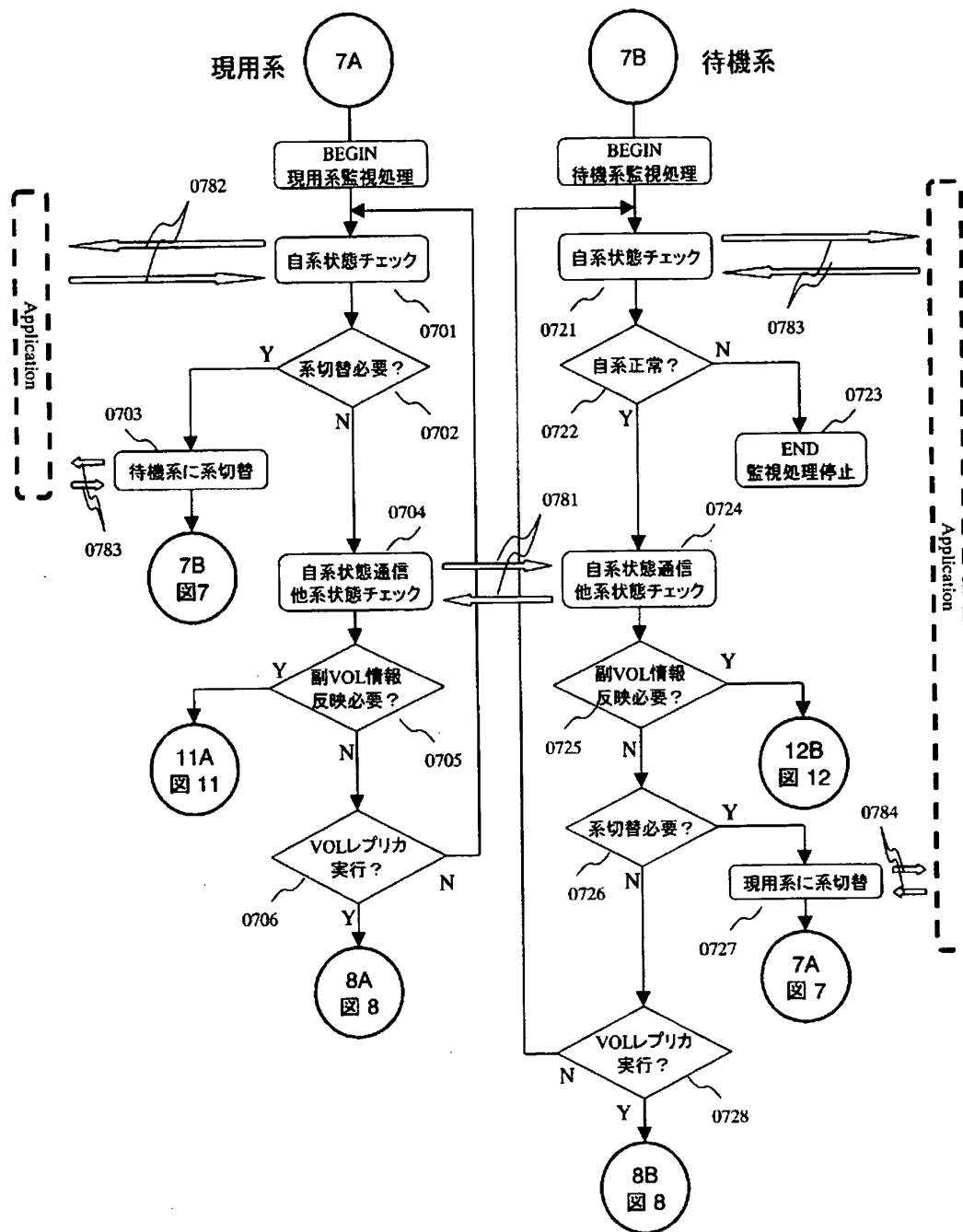
【図 6】

図 6

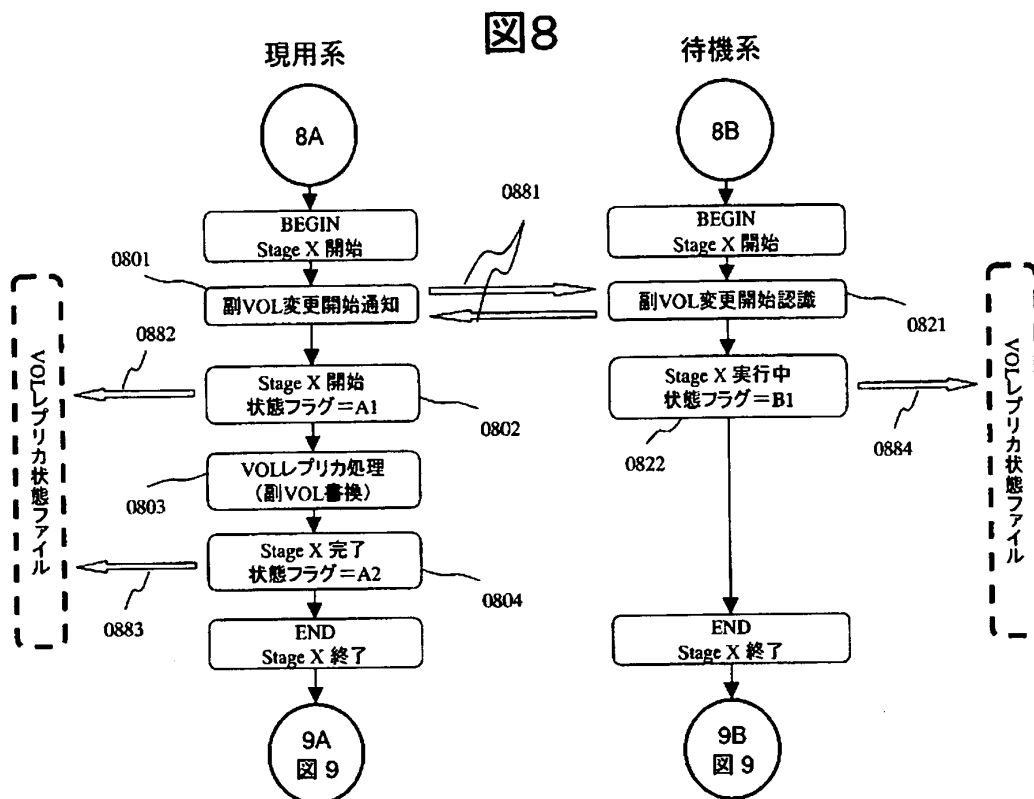


【図7】

図7

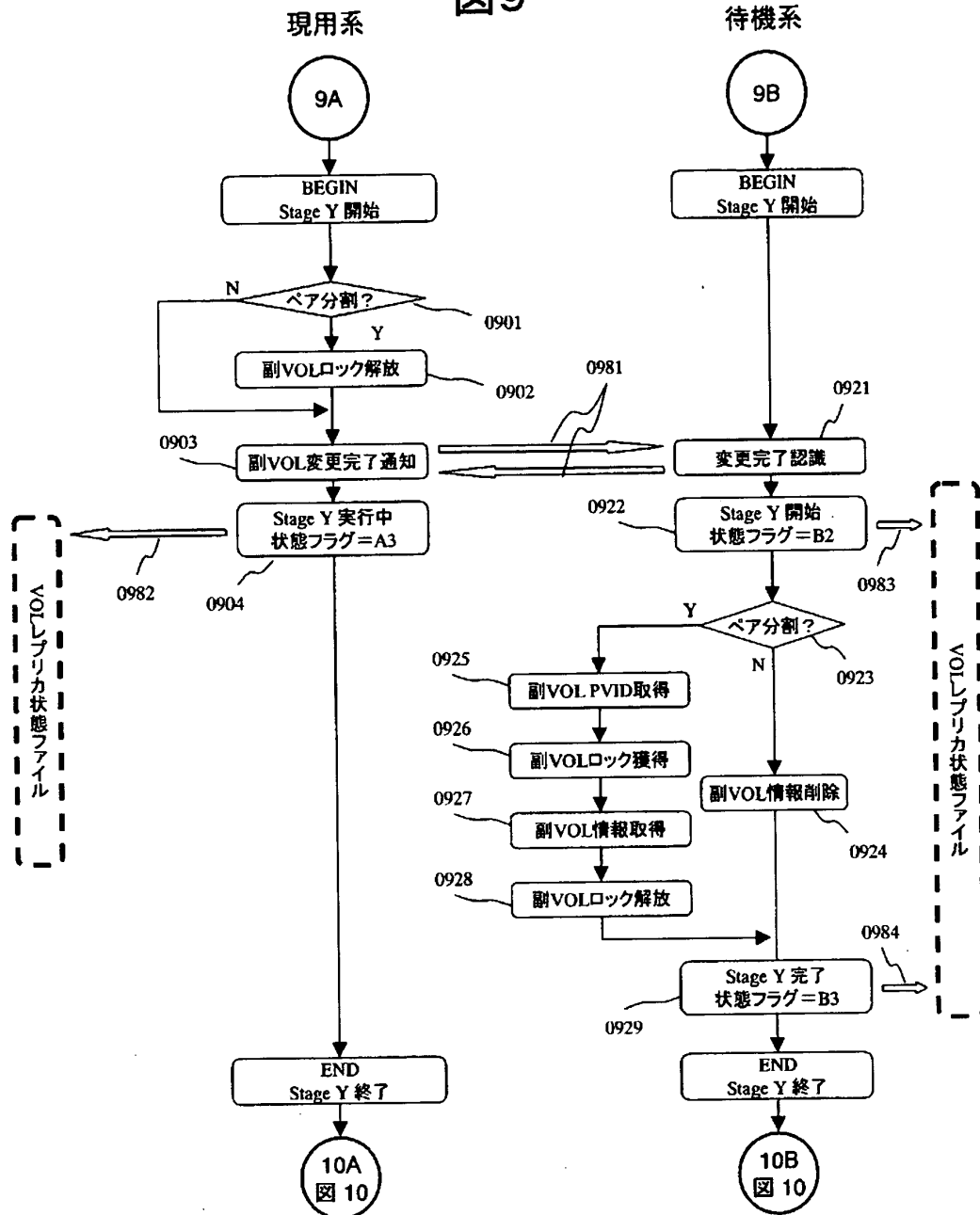


【図 8】



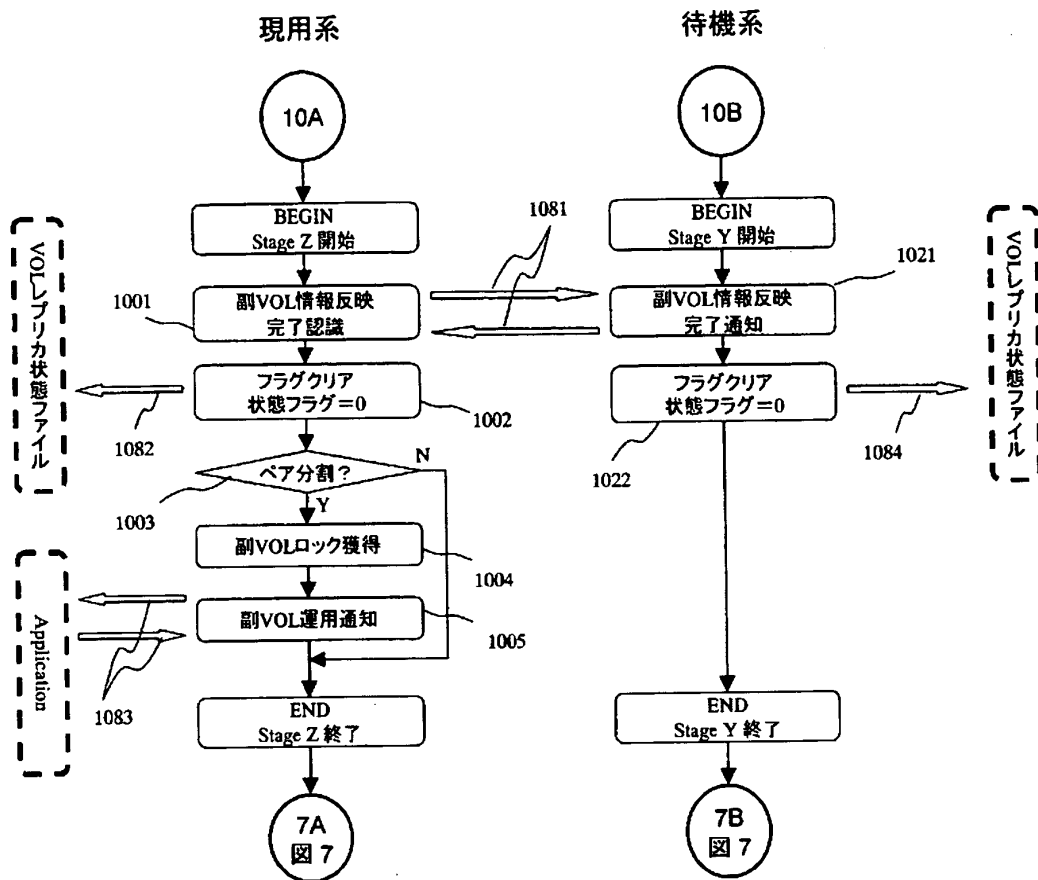
【図 9】

図 9

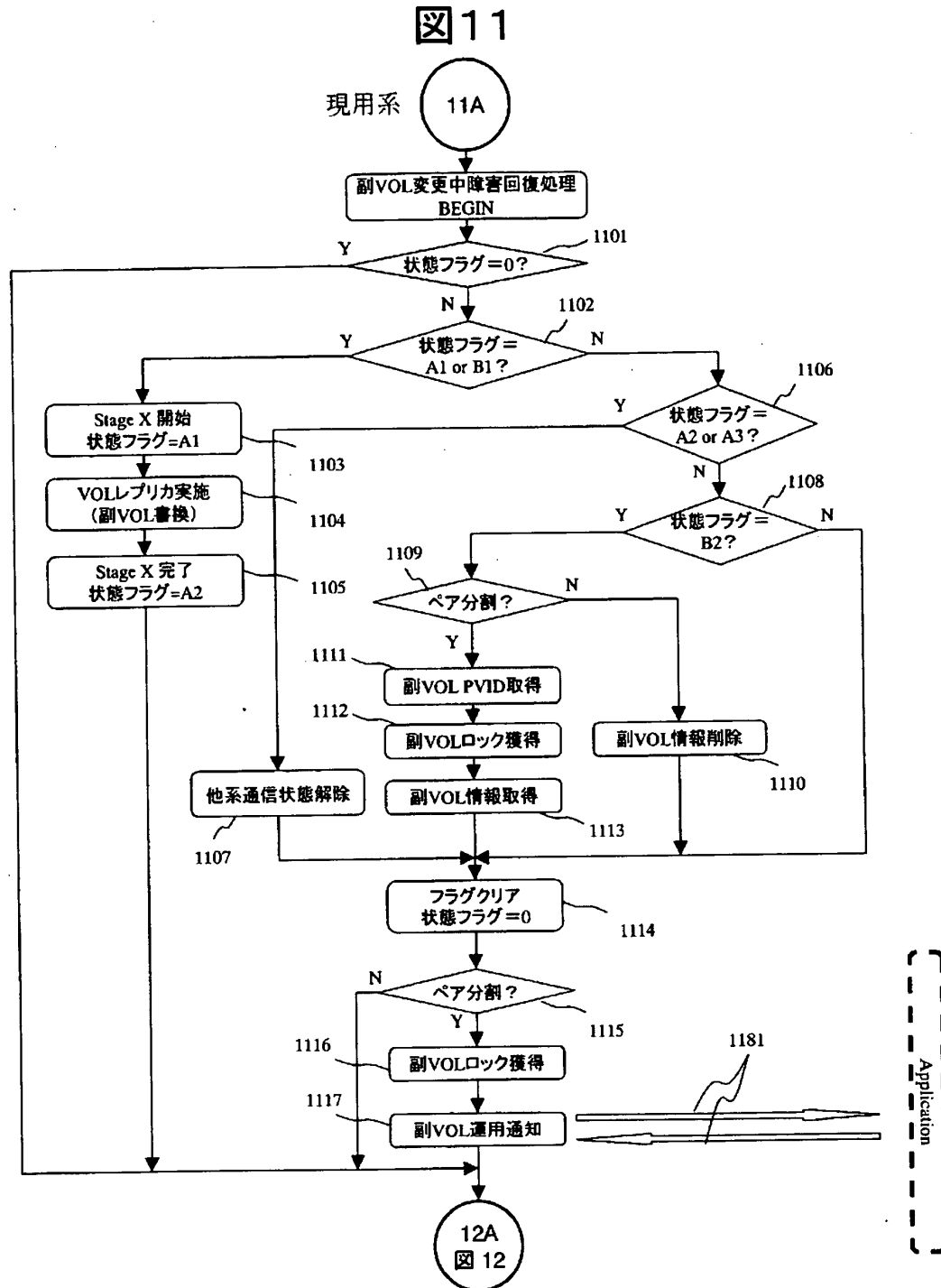


【図 10】

図10

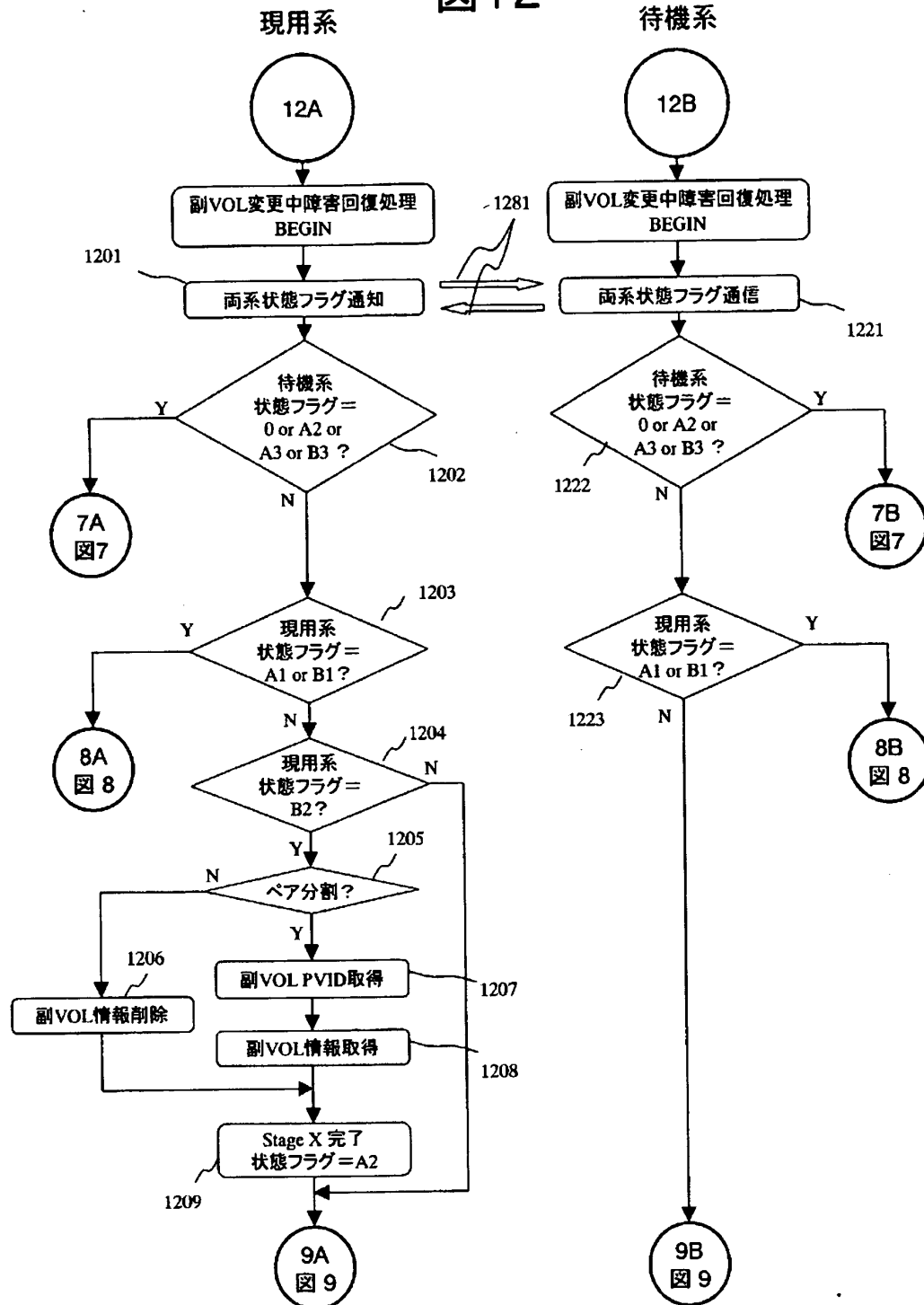


【図 11】



【図 12】

図12



【書類名】 要約書

【要約】

【課題】

VOLレプリカ手段による副VOLの変更を待機系に反映する方法及びシステムを提供することにある。

【解決手段】

ボリュームレプリカを用いて正ボリュームと副ボリュームを操作する現用／待機系コンピュータシステムにおいて、現用系がボリュームレプリカを実行後、副ボリュームが変更されたことを待機系に通知し、待機系は変更された副ボリューム情報を自系に反映する。これにより、現用系に障害が発生し、系切替が行われた後も、待機系反映された情報を元に副ボリュームに対してアクセスし、処理を続行する。

【選択図】 図 1

認定・付加情報

特許出願の番号	特願2003-057937
受付番号	50300353244
書類名	特許願
担当官	第七担当上席 0096
作成日	平成15年 3月 6日

<認定情報・付加情報>

【提出日】 平成15年 3月 5日

次頁無

特願 2 0 0 3 - 0 5 7 9 3 7

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所